DOCUMENT RESUME

ED 365 730 TM 020 970

AUTHOR Johnson, Colleen Cook
TITLE The Effects of Single and Compound Violations of Data Set Assumptions when Using the Oneway, Fixed Effects Analysis of Variance and the One Concomitant Analysis of Covariance Statistical Models.
PUB DATE Nov 93
NOTE 31p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 10-12, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; *Analysis of Variance; Computer Simulation; *Models; Monte Carlo Methods; *Robustness (Statistics); *Sample Size; Sampling
IDENTIFIERS *Balanced Designs; Unbalanced Designs; Variance (Statistical); *Violation of Assumptions

ABSTRACT

This study integrates into one comprehensive Monte Carlo simulation a vast array of previously defined and substantively interrelated research studies of the robustness of analysis of variance (ANOVA) and analysis of covariance (ANCOVA) statistical procedures. Three sets of balanced ANOVA and ANCOVA designs (group sizes of 15, 30, and 45) and one set of unbalanced ANOVA and ANCOVA designs (groups of 15, 30, and 45) were simulated. Violations of normal shape and three degrees of heterogeneity of group variances were included. Each set of ANCOVA analyses included ANOVA violations as well. Each data set violation was simulated in isolation and in combination, resulting in 665 unique ANOVA or ANCOVA F sampling distributions. Unbalanced designs almost always produced statistically invalid F ratios in the presence of any data set assumption except those that only involved perturbation of shape. In balanced designs, a degree of robustness beyond that expected was found, even when heterogeneity of variance was coupled with heterogeneous regression slopes and a skewed covariate. Robustness was found even when the ratio of largest to smallest group variance was five. Implications of these findings for educational, social science, and behavioral research are discussed. Three tables are included. (Contains 29 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

D

ED 365 730

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Colleen Cook Johnson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE EFFECTS OF SINGLE AND COMPOUND VIOLATIONS OF DATA SET ASSUMPTIONS WHEN USING THE ONEWAY, FIXED EFFECTS ANALYSIS OF VARIANCE AND THE ONE CONCOMITANT ANALYSIS OF COVARIANCE STATISTICAL MODELS

Colleen Cook Johnson

Presented at the Mid-South Educational Research Association Annual Meeting

November 12, 1993

TM020970

ABSTRACT

Capitalizing on the rapid advances of mainframe computer technology to date, this study integrates into one comprehensive Monte Carlo simulation a vast array of previously defined and substansively interrelated research studies spanning seven decades of methodological inquiry into the robustness of the analysis of variance (ANOVA) and analysis of covariance (ANCOVA) statistical procedures. Three sets of balanced ANOVA and ANCOVA designs (using equal group sizes of 15, 30 or 45) and one set of unbalanced ANOVA and ANCOVA designs (where the first group had an n of 15, the second group 30 and the third group 45) were simulated. Within each set of ANOVA analyses, violations of normal shape (including both skewed and non-mesokurtic dependent vectors) were included, as was three degrees of heterogeneity of group variances. Each set of ANCOVA analyses included the ANOVA violations, as well as those perturbations unique to ANCOVA: violations of homogeneity of group regression slopes and violation of the assumption that the covariate be normally distributed. Each data set violation was simulated both in isolation and in combination with one or more of the others; in the end resulting in 665 unique ANOVA or ANCOVA F sampling distributions. A modified version of the traditional methodology was developed and implemented - one which allowed for the systematic control of extraneous variables that traditionally have confounded the results of previous simulations.

As has been the case in previous research, the unbalanced designs almost always produced statistically invalid F ratios in the presence of any data set assumption except some situations that involveed only the perturbation of shape. When robustness was tested in the balanced ANOVA and ANCOVA situations, however, a degree of robustness far beyond that suggested by Glass, Peckham and Sanders (1972) was found, even in most situations where heterogeneity of variance was coupled with heterogeneous regression slopes and a skewed covariate. Robustness was found even when the ratio of the largest to smallest group variance was five.

This finding is particularly important to researchers in education, the social sciences and the behavioral sciences in light of the fact that the most common data set violation in the balanced design is heterogeneity of group variances. If a researcher can ascertain that their dependent variable's skew and kurtosis falls within the appropriate 95% confidence bands, then the ratio between the highest and lowest group variances can be as high as five without jeopardizing the robustness of ANOVA or ANCOVA results.

INTRODUCTION

Within a scientific discipline, theories unify the existing knowledge base as well as provide hypotheses for further extension of that knowledge base. Theories are abstractions, and as such are represented by the construction of conceptual or mathematical models; models which serve to abstract the subject under study while preserving the original structure of the system. By abstracting the subject into a succinct, parsimonious model, it is possible to determine how changes in one (or more) parts of a model might affect the system as a whole. Oftentimes these changes are impossible to observe and document in the real world, yet by manipulation of the model it is possible to shed light on both the effects of such change and the functioning of the model itself.

There are two types of models that can be developed: deterministic and probabilistic. Deterministic models are defined so that virtually 100% of the variance in the dependent variable can be explained by the independent variable(s) included in the model. For instance, "$E=mc^2$" can be considered a deterministic model if it can produce accurate estimates of E with little or no error. These models are seldom used in education, psychology or the social sciences. Concerning this, Lord and Novick (1968) write "deterministic models have found only limited use in psychology... because for problems of any real interest... we are unable to write an equation such that the residual variation in the dependent variable is small." (p. 23).

Instead, probabilistic models are more common in these disciplines. These models are not powerful enough to eliminate unexplained variation, although strategic methods are often used to minimize the proportion of unexplained variance while maximizing the amount explained. The general linear model (GLM) is a classic example of a probabilistic model. It has been argued that use of one specific form of the GLM, the analysis of variance, is the most widely used statistical procedure in education and the social sciences (Halpin and Halpin, 1988). Like other statistical models, those who use the GLM must assume that the prerequisite conditions for using the model actually do exist within their data set. However, a researcher seldom stumbles into a situation where all prerequisite conditions are perfectly met. Therefore, it is necessary to examine the statistical model itself, in its various forms, to determine to what extent real world conditions may depart from the assumptions inherent in the model before the GLM should be abandoned in favor of other statistical models.

The Nature of Monte Carlo Experimentation

There are two different kinds of mathematical research: theoretical and experimental. The main concern of theoretical mathematics is abstraction and generality. The theoretical

mathematician will write arguments in the form of symbolic expressions or formal equations which will abstract the essence of a problem, thus revealing the underlying structure. However, this strength is also its inherent weakness: the more general and formal the language, the less able the theory is at providing a numerical solution to a specific situation (Hammersley and Handscomb, 1967). The Monte Carlo approach allows the exploitation of the strengths of theoretical mathematics while avoiding the weaknesses inherent in it. Using this approach, experimentation replaces theoretical exploration when the latter falters.

Using Monte Carlo Simulation to Explore the Robustness of the General Linear Model

The GLM possesses a number of different forms, all of which provide an abstracted and succinct statement of the relationship between variables carefully chosen by research practitioners to reflect real world phenomena. Though the model is frequently used, the data collected for analysis never perfectly adheres to all of the assumptions of the model. Thus it becomes a question of how much difference there is between the conditions the model was designed to handle and the actual conditions that exist in a particular research situation. If the difference is within a "tolerable range," then use of one of the forms of the GLM should produce information that is statistically robust in its treatment of the relationship between variables. It is only when the data collected exceeds that "tolerable range" that alternatives to the GLM must be considered. Theoretical mathematics can be used to define the general nature of the problems that emerge when the GLM is used inappropriately, however it is unable to provide us with the precise limits of this "tolerable range."

Monte Carlo simulation provides valuable supplementary information about the problems that develop when assumptions underlying the GLM are violated. Using this methodology, it is possible to numerically define the degree of tolerance (i.e., robustness) that specific forms of the GLM have under real world research conditions.

This research is an empirical study of the effects of violations of the assumptions for two specific forms of the general linear model: the oneway, fixed-effects analysis of variance and the analysis of covariance using one independent variable and one concomitant (i.e., covariate). These two methods are used extensively in educational and psychological research, and serve as the mathematical foundation for more complex extensions of the GLM as well.

Unique Contributions of this Research

This study offers three unique contributions to the existing literature studying the appropriate use of ANOVA and ANCOVA in educational and psychological research. First, this study directly tested Harwell, Hayes, Olds and Rubinstein's 1990 claim that inflated type I error rates result when the ratio of largest to smallest group variances in the balanced design is as small

as two against the standard established by Glass, Peckham and Sanders (1972) that in balanced designs, one need only be concerned about the effects of heterogeneity of variances if the ratio of largest to smallest variances is at least three. Second, the study combined a number of different violations both separately and in combination, thereby examining the effects of data set violations at the zero, first, second, third and fourth orders. Most previous studies have been limited to exploration at the zero and first orders only. Finally, this study allowed for the systematic control of random noise that has confounded the results of past studies - thus providing findings that are more precise than those found in previous simulations.

REVIEW OF THE LITERATURE

The simplest prototype of the general linear model (GLM) is the t-test for two independent samples, which tests for mean differences between two groups. The oneway, fixed-effects analysis of variance (ANOVA) is the logical extension of this t-test; broadened in form to allow for the analysis of two or more groups. Both of these statistical procedures involve analysis of the effects of one discrete independent variable on a single, continuous dependent. In the oneway ANOVA, F represents the ratio of the variance in the dependent variable that can be explained by the researcher's data to that variance left unexplained. The analysis of covariance (ANCOVA) is a logical extension of the oneway ANOVA, applicable when a third, continuous variable (refered to as the covariate or concomitant variable) is known to have a significant effect on the dependent variable, while having little or no effect on the independent variable. When ANCOVA is appropriate, the researcher's goal is to probe the effects of the independent variable on the dependent after removing the influence of the concomitant. To do this, ANCOVA first removes all variation in the dependent variable that is a function of the concomitant. Then, using these "adjusted scores," ANCOVA effectively reanalyzes the data for mean differences between the groups that make up the independent variable.

The two forms of the GLM studied in this simulation are the oneway, fixed-effects ANOVA and the one concomitant ANCOVA. Most researchers in this area accept the premise proposed by Cochran (1957) and Winer (1962), who assert that the relationships found in the basic oneway ANOVA and even the more basic t-test for two independent samples carry over into the ANCOVA extension. Therefore, this literature review will contain discussion of relevant theoretical and empirical research involving the use of all of these statistical models.

Statistical Models and the Assumptions Inherent Within Them

When they are initially developed, statistical models (ie, procedures) are designed to be used under a specified set of conditions (that is, assumptions about the data set that the model is

used to describe). These conditions are designed to balance creditability (the ability to process data in a form that will be useful to researchers) with manageability (the technique's ability to simplify many mathematical derivations and operations). If valid results are to be obtained, the researcher must assume that his or her data set is similar to the type of data set required by the statistical procedure chosen.

Seldom, however, do data sets adhere perfectly to the assumptions a statistical model was developed to handle. According to Glass, Peckham and Sanders (1972), the question that the researcher must ask in reference to the data collected is not *whether* the assumptions are satisfied, but instead, *are the violations that do occur extreme enough to compromise the validity of the results?*

Box and Anderson (1955) argue that to fulfill the needs of the researcher, statistical criteria should: (1) be sensitive to change in the specific factors being tested (in other words, they should be powerful) and (2) they should be insensitive to changes in extraneous factors of a magnitude likely to occur in practice (in other words, they should be robust).

Literature Concerning the Assumptions of the Oneway, Fixed-Effects ANOVA

In 1972, Glass et al. identified three assumptions of concern for the ANOVA. The first of these is additivity - that is, each observation must be the simple sum of three components: the grand mean (μ), the effects of the treatment (a_j) and the error associated with the individual observation (e_{ij}). The presence of additivity is important because the least amount of information is lost in an additive model (Cochran, 1947). The second assumption is that the sum of the treatment effects equal zero. Glass et al. argue that this assumption is actually a mathematical restriction adopted to allow for a unique solution to the least-squares equation, rather than an assumption per se. Finally, the third assumption is that errors made while using the model should be normally distributed with a population mean of zero and a variance of ^{2}. This third assumption involves the nature of the errors in the population from which the data originates, and takes three distinctive forms: (a) normality of the error distribution, (b) homogeneity of group variances, and (c) the independence of errors. Independence of errors is, of course, a methodological concern. Therefore it is forms (a) and (b) of the third assumption that are the subject of most theoretical and empirical research into ANOVA.

Homogeneity of Variance

This assumption was first identified in the classical 1908 paper "The Probable Error of the Mean" by The Student (Gossett), however, the publishing of empirical results in this area would wait until the work of Hsu (1938, as cited by Scheffe', 1959). Active research concerning the assumption of homogeneity of variance has continued even until today. Many of the published

studies suggest that the F test with equal sample n's is robust when faced with the single violation of the assumption of unequal group variances as long as the ratio of the largest to smallest group variances does not exceed three (e.g. Glass et al, 1972) Some studies (e.g. Shields, 1978) suggest that the degree of robustness present may be offset by the loss of power that is the result of using a parametric test when heterogeneity of group variances is present. The validity of the F ratio, however, is questionable in situations where both the sample sizes and variances are unequal. When cell sizes are unequal and two groups are involved, research suggests that inflated type I error rates occur when the larger group size is paired with the smaller group variance (e.g., Scheffe´, 1959). But the most surprising results of recent years, however, came in a meta-analytic study conducted by Harwell, Hayes, Olds and Rubinstein. They suggest that even when sample n's are equal, inflated type I errors are possible when the ratio of largest to smallest variance is as small as two. Thus, Harwell et al. (1990) write,"... researchers should not rely on equal sample sizes to neutralize the effects of heterogeneous variances" (p. 23).

Normality of the Distribution of Errors

Research dating back to the 1920's has investigated violations of this assumption. Games and Lucas (1966) suggest that skewed distributions are a greater threat to robustness than leptokurtic or platykurtic distributions, however this claim is not consistent with Pearson's 1929 power analysis among balanced designs. Assuming a distribution with a mean of zero and variance of one, the third moment (from which skewness is mathematically derived) is defined as follows:

$$b_1 = \frac{\Sigma(X - m)^3}{3}$$

while the fourth moment (used to calculate kurtosis) is defined as:

$$b_2 = \frac{\Sigma(X - m)^4}{4}$$

Norton (1952, cited in Glass et al., 1972) examined the degree of skewness in data distributions and found a moderately skewed distribution as having a skew value around .5, while the skew value for an extremely skewed distribution was around 1.0. A perfectly symmetrical distribution (in other words, a distribution with no skew) has a skew of 0. A perfectly mesokurtic distribution has a kurtosis of 0. Distributions with kurtosis significantly greater than zero are leptokurtic, while those significantly less than zero are platykurtic.

Looking at the effects of skewness in the single sample t-test, Pearson (1929) and Scheffe´ (1959) found that if the difference between the sample and population mean is positive and the distribution is positively skewed, then actual power will exceed nominal power. However,

if the difference between the sample and population means is positive and the distribution is negatively skewed, then the actual power is less than nominal power. Games and Lucas suggest that F-test results may improve when the procedure is conducted on data that has highly leptokurtic error distributions, while F test results for data with platykurtic error distributions tend to be adversely affected.

Extension of ANOVA Assumptions to ANCOVA

The simplest form of the analysis of covariance (which consists of one independent, one concomitant and one dependent variable) is an extension of the oneway, fixed-effects ANOVA. According to Cochran (1957) and Winer (1962), the assumptions previously discussed in regards to ANOVA apply to ANCOVA as well, provided that the concomitant variable is normal. It is for this reason that empirical testing of either of these single violations in the ANCOVA case is scarce.

The sensitivity of the F-test in ANCOVA to departures from normality in the dependent variable depends on the degree of nonnormality that is found in the concomitant (Potthoff, 1965). Similar results were found in Atiquallah's theoretical treatise (1964): if X (the concomitant) is a normally distributed random variable, nonnormality in the dependent variable has little effect on the F-test. If, however, the concomitant is a random variable that is not normally distributed, then there will appear an increased sensitivity of the F-test to nonnormality in the dependent variable.

The Seven Assumptions of the Analysis of Covariance

Elashoff (1969) and McLean (1979, 1989) report the following seven assumptions associated with ANCOVA: (1) the cases are assigned at random to treatment conditions; (2) the covariate is measured error-free (that is, there is a perfect reliability in the measurement of the covariate); (3) the covariate is independent of the treatment effect; (4) the covariate has a high correlation with the dependent variable; (5) the regression of the dependent variable on the covariate is the same for each treatment group; (6) for each level of the covariate, the dependent variable is normally distributed; and (7) the variance of the dependent variable at each given value of the covariate is constant across treatment groups. These assumptions can be classified as falling into one of two categories: (a) assumptions that are concerned with the research design and sampling (methodological assumptions) and (b) assumptions that are concerned with the numerical form of the data set and the population from which it came (data set assumptions).

Methodological Assumptions

Two of the ANCOVA assumptions deal with the research design and sampling: (1) the cases are assigned to random treatments (randomization) and (2) the covariate has perfect reliability. Concerning the issue of randomization, Evans and Anastasio (1968) distinguish three

separate situations: (1) individuals are assigned to groups at random after which the treatments are randomly assigned to the groups; (2) intact groups are used, but treatments are randomly assigned to the groups; and (3) intact groups are used where treatments occur naturally rather than being randomly assigned by the researcher. They maintain that ANCOVA is appropriate for the first situation, can be used with caution in the second, but should be abandoned altogether (perhaps in favor of the less restraining factorial block ANOVA design) in the third. Two reasons are provided for their recommendations: first, it is never quite clear whether the covariance adjustment has removed all of the bias when proper randomization has not taken place, and second, when there are real differences among the groups, covariance adjustments may involve computational extrapolation.

Raajimakers and Pieters (1987) and also McLean (1974) have addressed the issue of an unreliable covariate. Raajimakers and Pieters note that there are two ways that the researcher can conceptualize covariate reliability. If one assumes that the dependent variable is linearly related to the observed value of the covariate, then the ANCOVA results will retain their statistical validity. If, on the other hand, it is assumed that the dependent variable is linearly related to the underlying true score on the covariate (rather than the sample of scores that were actually observed), then the resulting F ratio will produce biased results. McLean's research, however, suggests that the issue of perfect reliability becomes less of a threat to the validity of the F ratio if there is an independence of the covariate measure and the treatment groups.

The Covariate's Relationship with the Independent and Dependent Variables

The covariate should have no significant correlation with the independent variable, yet be highly correlated with the dependent variable. Feldt (1958) recommends the use of a covariate only when the zero-order correlation between the covariate and the dependent variable is $r \geq 0.6$. McLean (1979, 1989) sees the relationship between the covariate and the independent variable to be the most fundamental of all of the assumptions, and suggests that ANCOVA not be performed until after the data has been tested to see if it meets this assumption. If this assumption is not met, the F-test results are not invalidated as such, however it reduces the ANCOVA's efficiency to slightly below that of doing a simple oneway ANOVA on the same data.

Homogeneity of Group Regression Slopes

This assumption requires that the slope of the regression line between the concomitant and dependent variables be the same for all levels of the grouping variable. The problem, if this assumption is violated, is analogous to trying to interpret main effects in the presence of significant interactions in an n-way factorial ANOVA. If heterogeneous regression slopes are

suspect, the researcher would be wiser to use the randomized block ANOVA rather than ANCOVA.

Peckham (1968), McClaren (1972) and Hamilton (1972) have investigated the effects of violation of this assumption. Peckham varied regression slopes, the number of groups, and the sample size, though he limited himself to equal groups. Values of the concomitant variable were fixed and chosen to conform as closely as possible to a normal research situation. He found that there were small discrepancies in the actual vs. theoretical significance levels when the slopes were varied. He also found that as the degree of heterogeneity of the regression slopes increased, the heterogeneity of group variances likewise increased, and therefore the empirical rate of the Type I errors decreased from what is suggested by normal theory.

McClaren found similar results to Peckham when he looked at equal samples; however he extended his study to unequal groups. With the unequal group n's, McClaren found results similar to those reported by Box (1954) and Scheffe´ (1959); that is, when the smallest regression coefficient and the largest variance were combined with the smallest sample size, the empirical significance levels were biased in a non-conservative direction, and likewise, when the pairings were reversed, the test became conservative.

When Hamilton conducted his study, he limited his analysis to two groups. He used the same combination of equal sample sizes, number of groups, and regression coefficients as Peckham and McClaren, yet failed to replicate their findings. Whereas Peckham and McClaren observed a conservative bias in empirical alpha levels when sample n's and regression slopes were heterogeneous, Hamilton's values were close to nominal alpha. It is unclear why there is a discrepancy in the results of the three studies (Shields, 1978). Theoretical work by Atiquallah (1964), however, suggests that ANCOVA should be robust enough to the violation of the single assumption of homogeneity of regression in situations where the sample size is large and the means of the concomitant variable(s) are equal. Otherwise, Atiquallah suggests, the test should be biased in a conservative direction.

Homogeneity of Variances and Nonnormal Error Distributions in ANCOVA

As has been discussed previously, most researchers simply accept the claim by Cochran (1957) and Winer (1962) that the effects of the simple ANOVA violations are equally viable when the model is extended to include one or more concomitant variables.

RESEARCH METHODOLOGY

Goal of the Research

This research is an exploratory study of the effects of both single and compound violations of the mathematical conditions (i.e., assumptions) underlying use of the analysis of

variance and covariance designs. Monte Carlo methodology was used, allowing for the empirical investigation of problems identified by theoretical mathematicians as potential threats to the robustness of the ANOVA and/or ANCOVA results under conditions common to research practitioners in the behavioral sciences, social sciences and education. Because of advances both in methodological techniques and computing technology, the capability has emerged to study this topic in depth, yet with a global perspective not possible just a few years ago. Capitalizing on these advances, this study has integrated into one comprehensive laboratory experiment a vast array of previously defined and substantively interrelated research avenues that have spanned across seven decades of statistical inquiry.

Specifically, this research explores the following violations that can occur in a researcher's data set: heterogeneity of group variances, skewness, non-mesokurtic distributions, and (in ANCOVA) heterogeneity of regression slopes and use of a skewed concomitant.

Information about the Computing Environment and the Programs Written to Conduct the Simulations

The statistical simulations were conducted on a Digital Equipment Corporation VAX 6430 mainframe computer with 128 M-bytes of MOS memory and 32 gigabytes of disk storage space. The simulation itself consisted of two sets of eight FORTRAN 77 programs written especially for this research: the first set of programs (phase one of the simulation) conducted simulations that used a normally distributed covariate vector, while the second set of programs (phase two of the simulation) conducted the same analyses using a skewed covariate vector. The data generated by the experiments in phase one were used again in phase two with one exception: the concomitant vectors generated for phase one were mathematically perturbed to produce the skewed concomitant vectors needed for phase two.

The Simulation Process, Part I: Within a Single Replication of an Experiment

Four experimental situations were simulated in each of the two phases of the simulation: three balanced designs (i.e., equal sample sizes) and one unbalanced design (i.e., unequal sample sizes). For explanation purposes, these four experimental situations will be refered to in this text as experiments A, B, C and D. Experiment A tested the ANOVA and ANCOVA F statistic when three equal groups of size 15 were used. Experiment B involved simulation using three equal groups of size 30, while experiment C tested the F statistic when three equal groups of size 45 were used. The fourth condition, experiment D, involved simulation of the ANOVA and ANCOVA F statistic when three unequal sized groups (n's = 15, 30 and 45) were used.

Experiments A, B and C of phase one were used to generate the data. Experiment D, on the other hand, did not generate data. Instead it imported grouping, concomitant and dependent

vectors from the data generating experiments, so that the first group had a size of 15, the second group 30 and the third group 45. The use of data in experiment D which was not independent of the data used in experiments A, B and C was to facilitate the comparison of the balanced and unbalanced design results. By using the same data, a major source of sampling error was eliminated; sampling error that otherwise might confound interpretation of the results. Likewise, phase two of the study (for both the balanced and unbalanced designs) imported data that was created in the data generating experiments of phase one with only one change: the concomitant vectors, which were normally distributed when they were originally created in phase one, were perturbed to create a moderately skewed covariate.

The data generated for experiments A, B and C were created using the International Mathematical and Statistical Libraries (IMSL) subroutine RNVMN, a subroutine which is designed to create multivariate normal distributions with means equal to zero, standard deviations equal to one, and correlations between vectors that can be specified beforehand by the user. Data for each treatment level were created separately using IMSL. This made it possible to obtain the unequal group regression slopes desired for the second concomitant vector. For the first concomitant vector, the correlation between all groups and the IMSL created dependent variable was set at $r = 0.707$, thus simulating homogeneity of regression slopes. For the second concomitant, heterogeneity of regression slopes was simulated by having IMSL create concomitant vectors for group 1 that had a correlation of $r = 0.6$ with group 1's dependent vector, a correlation of $r = 0.707$ between the group 2 concomitant and dependent vectors, and $r = 0.8$ between the third group's concomitant and dependent vectors.

The next step of the data creation process would require that duplicate copies of the dependent vector be created and then perturbed in a systematic fashion to simulate specific skew and/or kurtotic conditions. Therefore, it was imperitive that the originally created vectors themselves have the proported mean, variance, skewness and kurtosis. This was accomplished by building a testing procedure into the data generating FORTRAN programs.

By using this testing procedure, dependent vectors created by IMSL were tested to see if their skew and kurtotic values fell within the 95% confidence bands that surround zero skew and kurtosis for the specific group size. Therefore, for experiment A (where the group size was 15), all dependent vectors generated by IMSL were tested to determine if their skew was between -1.137 and 1.137, while the kurtosis was tested to see if it fell between -4.038 and 4.038. If either value was not within these limits, then the data created by IMSL was discarded, and new data created and tested. Likewise for experiment B ($n = 30$), skew values were tested to assure that they fell between the values of -0.837 and 0.837, while kurtosis values were checked to assure that they were between -3.478 and 3.478. For experiment C, confidence bands for skew were -0.693 and 0.693, while they were -3.205 and 3.205 for kurtosis. For all of the data generating

experiments, the data was retained only when both the skew and kurtosis values of the dependent vectors created by IMSL fell within these limits.

These checks assured that the base vectors, (that is, those created originally by IMSL) were normally distributed, with no significant skew or kurtosis. This, in turn, allowed for mathematically valid perturbations to be performed on them. The checks do, however, represent a departure from the sampling procedure characteristic of more traditional Monte Carlo studies. Using the more traditional approach, parent populations with the desired mathematical characteristics are created. Out of these parent populations, repeated samples of the desired size are randomly selected and tested. While this methodology is more generalizable because of its ability to simulate the central limit theorem, it also allows the inclusion of samples with skew and/or kurtosis radically different from what they are proported to be. Therefore, when differences between the empirical results and normal theory surface, it is unclear to what degree these differences are the result of the known mathematical characteristics of the parent population, and at what point they become the result of selected samples that, as the result of pure chance, possess mathematical characteristics far different from their parent population.

After IMSL created acceptable concomitant and dependent vectors, phase one of the simulation required that the normal dependent vector be duplicated then algebraically perturbed to simulate 27 different mathematical conditions. Distortions of distributional shape were imposed on the data first. This was done using Fleishman's method (1978), which uses the following function:

$$Y = a + bX + cX^2 + dX^3$$

where the coefficients b, c and d are obtained by consulting a special table compiled by Fleishman, and the coefficient a has the same absolute value as the coefficient c, but the opposite sign. Using this polynomial expression, the base dependent vector's values were substituted for X, while the resulting Y values formed a distribution with the desired shape.

Use of Fleishman's function allowed the desired combination of skew and kurtosis values to be created within a tolerable margin of error without distorting the original mean or standard deviation. The originally created (ie., base) dependent vector was normal, with no skew and kurtosis. After Fleishman's formula was imposed on duplicate copies of the original dependent vector, the following combinations of skew and kurtosis were simulated: moderately skew (skew = 0.5, kurtosis = 0), platykurtic (skew = 0, kurtosis = -0.5), leptokurtic (skew = 0, kurtosis = 2), moderately skewed and platykurtic (skew = 0.5, kurtosis = -0.5), moderately skewed and leptokurtic (skew = 0.5, kurtosis = 2), and extremely skewed and leptokurtic (skew = 1, kurtosis = 2). This allowed for every combination of skew and kurtosis with two exceptions: an extremely

skewed and platykurtic distribution and an extremely skewed and mesokurtic distribution. Neither of these shapes were possible to obtain using the coefficients published by Fleishman (1978).

After the algebraic manipulations to distort shape, seven dependent vectors possessing the characteristics described above were available. Each of these seven vectors were then duplicated three more times, and the three duplicate vectors for each shape linearly transformed. Afteer the duplicate vectors were transformed, there were four different group variance ratios for each of the seven distributional shapes: 1:1:1 (homogeneity of variance), 1:1.5:2 (slight heterogeneity of variance), 1:2:3 (moderate heterogeneity of variance) and 1:3:5 (extreme heterogeneity of variance). These inter-group variance conditions were chosen specifically to allow the testing of Harwell et al.'s 1990 claim (that differences from normal theory may be present in balanced designs when the ratio between the largerst and smallest variance is 2) against the standard set by Glass et al. (that differences from normal theory do not emerge in balanced designs until the ratio between the largest and smallest variance is at least 3).

As has been mentioned previously, no new data was generated for experiment D (the unbalanced design). Instead, a systematic process imported vectors already created. Specifically, treatment level (group) 1 from experiment A, group 2 from experiment B, and group 3 from experiment C were imported. This created the unequal n simulation where group 1 had an n = 15, group 2 had an n = 30, and group 3 had an n = 45.

Therefore, in the end 28 different dependent vectors, two concomitant vectors and a grouping vector were either created for or imported into each replication of all of the experiments. For the ANOVA simulations, the grouping vector was combined with each of the dependent vectors, computing 28 F ratios (one for each combination of skew, kurtosis and variance). For the ANCOVA simulations of phase one, the first concomitant vector was combined with the grouping vector and each of the 28 dependent vectors to calculate 28 ANCOVA F statistics using a normal covariate with equal regression slopes. The second concomitant vector was then combined with the grouping vector and each dependent vector to calculate 28 ANCOVA F statistics using a normal covariate with unequal regression slopes.

As has been mentioned before, the experiments of phase two used the same data that was created in phase one, however the normal covariate created in phase one was skewed by Fleishman's function (skew value = 0.75). Phase two was designed to test ANCOVA when the only difference was use of a skewed concomitant rather than a normal one. Therefore, only 56 additional F statistics were calculated per replication in this phase: 28 involving use of a skewed covariate and equal slopes and 28 involving use of a skewed covariate and unequal slopes.

Besides using IMSL subroutines to generate the data, IMSL subroutines were also incorporated into the FORTRAN programs to calculate the F ratios. Specifically, IMSL subroutine

AONEW was used to obtain the ANOVA F values, while subroutine AONEC was used to calculate the ANCOVA F values.

The Simulation Process, Part II: The Global Design

As has been mentioned previously, phase one of the study was designed to test the ANOVA and ANCOVA F test when a normal covariate was combined with violations of one or more of the following assumptions: normal skew, normal kurtosis, homogeneity of variances and (in the ANCOVA) situation, homogeneity of regression slopes. Phase two of the study conducted the same analyses using a skewed covariate rather than a normal one.

Glass et al. (1972) recommended that the sampling distributions created in Monte Carlo studies have a minimum of 2000 F ratios each. For the three experimental conditions involving equal group n's, sampling distributions of 4000 (twice the minimum recommended by Glass et al.) were created. In the experimental condition involving unequal n's and homogeneity of variances, F sampling distributions of 4000 F ratios were also created. In the situation where unequal n's were combined with heterogeneity of variances, however, the combination of variance ratios and group sizes were varied so that two sets of sampling distributions with 2000 F ratios each were developed: one set where the largest group variance was combined with the largest sample size and the other set where the largest group variance was combined with the smallest sample size. This was done since previous literature suggests that heterogeneity of group variances produces different effects in the unequal n situation, depending on the combination of sample size and magnitude of group variances (e.g. Box, 1954, McClaren, 1972 and Scheffe', 1959). The relationship between sample size and group regression coefficients was fixed for those analyses that involved unequal group slopes, therefore the process of varying the magnitude of group variances with the sample size produced the following triple combinations for analysis in the ANCOVA simulations: (1) the largest group size with largest group variance and largest regression coefficient, and (2) the largest group size with the smallest group variance and largest regression coefficient. Previous literature (e.g. Glass et al., 1972, Shields, 1976) suggest that the additivity of effects should produce dramatic differences in these two combinations.

After running all four sets of experiments in both phases of the simulation, a total of 420 empirical sampling distributions of 4000 F ratios each were created, representing all single and compound data set violations for the balanced ANOVA and ANCOVA simulations. Another 35 sampling distributions of 4000 F ratios each included all unbalanced ANOVA and ANCOVA simulations with homogeneity of group variances. Finally, another 210 empirical sampling distributions of 2000 F ratios each were created, representing all single and compound data set violations having both heterogeneous variances and unequal sizes.

Of these 665 F sampling distributions, four ANOVA and four ANCOVA F distributions were created using data that did not contain any violation under study. These eight sampling distributions (one ANOVA and one ANCOVA for each of the four experimental conditions A, B, C and D) served as a baseline against which other distributions could be compared, and served as a check to make sure that the simulation was operating properly.

Statistical Analysis of the Sampling Distributions

In addition to qualitative evaluation of the sampling distributions, statistical analysis of the data was performed using the Kolmogorov-Smirnov one sample test at the $p < .05$ and (where applicable) $p < .01$ levels of significance. The non-parametric test was employed to compare the empirical sampling distributions with the appropriate theoretical (i.e., nominal) F distribution at four key points in the nominal F tail region: .90, .95, .975 and .99. These points, of course, are the points on the nominal F curve used by practitioners when testing for significance at the $p < .10$, $p < .05$, $p < .025$ and $p < .01$ levels of significance respectively. In addition, the means, standard deviations, skew and kurtosis values for each of the entire populations of data generated in the study were calculated and inspected to assure the integrity of the results.

RESULTS

Summarized here are the specific results of the effects of violations of data set assumptions for the analysis of variance and covariance statistical models. Since the integrity of the results is dependent on the quality of the data produced, the first section will take a look at descriptive statistics for the entire population of data produced for this simulation. The second section will summarize the effects of violations in the ANOVA situation. The third and fourth sections will summarize the effects of violations on the ANCOVA.

Analysis of the Population Data

All data created in each of the replications of the data generating experiments were retained in order to verify the integrity of the results. In the actual process of creating the data, the vectors for each treatment level were created individually then merged with the vectors for the other treatment levels before ANOVA or ANCOVA could be performed. The population vectors were checked for each treatment level separately, then the full vectors (which consisted of the three treatment levels merged together) were also checked. All population vectors, including the base vectors created by IMSL and the vectors perturbed by use of Fleishman's function were at or very near their target parameters.

The size of the populations are worth noting. In their classic 1972 paper, Glass et al. suggest that populations with the desired characteristics have a minimum of 10,000 points each. The population N's used in this study were considerably larger than the minimum standard: 180,000 for each of the full population vectors created in experiment A, 360,000 for the full population vectors created in experiment B, and 540,000 for the full population vectors created in experiment C. The population statistics for the vectors created to simulate heterogeneous variances were also checked. As expected, the simple linear multiplication that changed their variances did not change the vectors means, skew or kurtosis.

Effects of Data Set Assumptions on the Analysis of Variance

For all of the analyses to follow, comparisons were made between the empirical F sampling distributions and the theoretical (i.e., nominal) F distributions expected using normal theory. In calculating the values included on these tables, the difference scores recorded on the table were found using simple subtraction: the number of observed F ratios minus the number expected using normal theory.

When the group size was 15 and all groups were equal, no empirical sampling distribution was found to have type I error rates significantly different from what would be expected under normal theory, although the sampling distribution that was based on an extremely skewed and leptokurtic dependent vector with extreme heterogeneous variances (variance ratio 1:3:5) came within one F value of being significant at the $p < .05$ level. When violations were imposed on the dependent vectors with groups of size 30 and 45, no empirical distributions were found significantly different from the nominal F distribution at the $p < .05$ level (See Table 1).

For the equal n experiments, the differences between the empirical and theoretical F sampling distributions were largest when the sample size was small and became smaller as the group sizes grew larger. It is possible that this trend, found in the ANCOVA results as well, may be due to the fact that confidence bands increase when sample size is small. All dependent base vectors created by IMSL, as one will recall, were tested to exclude extreme vectors with mathematical characteristics different from those proported. It is possible that when sample sizes are less than 30, confidence bands are not narrow enough to eliminate all samples that are not representative of their parent populations.

No significant differences were found in the unbalanced designs having homogeneous variances. Significant differences did emerge, however, when the unbalanced ANOVA was combined with even the slightest degree of heterogeneity of variance (group variances as small as 1:1.5:2). Further analysis revealed two different trends, depending on whether the largest variance was coupled with the largest or smallest group. Specifically, when the smallest group had the largest variance, all empirical sampling distributions were significantly less than the theoretical

Table 1

Maximum Differences Between the Empirical and Nominal Sampling Distributions for the ANOVA Simulations

Largest to Smallest Group Variance Ratios	BALANCED DESIGNS										UNBALANCED DESIGNS					
	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	2:1	3:1	5:1
	Sample Size = 15				Sample Size = 30				Sample Size = 45				Sample Sizes = 15, 30, 45			
Distributional Shape																
Normal	17	-15	-34	-60	17	-17	-24	-37	-9	-16	-17	-34	27	88¶	107¶	115¶	-87¶	-136¶	-194¶
Platykurtic	30	-8	-33	-63	21	-14	-23	-39	6	-12	-17	-22	26	87¶	109¶	119¶	-86¶	-142¶	-195¶
Leptokurtic	24	-11	-27	-52	-8	-28	-31	-44	-14	-11	-20	-35	16	85¶	102¶	113¶	-84¶	-129¶	-193¶
Moderate Skew	24	-13	-39	-73	9	-17	-28	-44	11	-11	-19	-31	28	83¶	103¶	112¶	-86¶	-144¶	-195¶
Mod. Skew & Platy.	27	-16	-39	-67	14	-10	-18	-35	11	-7	-13	-19	34	91¶	111¶	118¶	-84¶	-132¶	-201¶
Mod. Skew & Lepto.	15	-14	-27	-65	-6	-24	-34	-43	-16	-14	-20	-32	19	84¶	99¶	111¶	92¶	-136¶	-198¶
Extreme Skew & Lepto.	23	-17	-43	-85	2	-12	-30	-46	-16	-20	-25	-44	24	87¶	106¶	115¶	-96¶	-143¶	-200¶

¶ Significant at the $p < .01$ level

F distribution at the $p < .01$ level of significance. When the largest group contained the largest variance, however, the opposite trend developed: sampling distributions having heterogeneity of variances were found to be significantly greater than theoretical F at the $p < .01$ level.

Effects of Assumption Violations on the Analysis of Covariance Using a Normal Concomitant

Differences between the empirical and nominal F sampling distributions for the ANCOVA simulations using a normal covariate are found in Table 2. For the balanced design using small but equal group sizes ($n = 15$), the only compound violation that had a significant impact on the resulting empirical sampling distribution was the combination of an extremely skewed and leptokurtic shape with extreme heterogeneity of variances (ratio of 1:3:5), which was significant at the $p < .05$ level. In those simulations that had equal n's of size 30, no significant differences emerged. Equal n's of 45 showed more of the same; no significant differences were found even when extreme heterogeneity of variances was combined with unequal regression slopes.

Among the unbalanced ANCOVA simulations involving homogeneity of variances, no significant differences emerged as long as the regression slopes were equal. When the group slopes were unequal, however, all analyses were significant at the $p < .01$ level.

In those ANCOVA simulations involving both equal slopes and heterogeneous variances, significant differences emerged - most at the $p < .01$ level. Different trends emerged, however, depending on whether the largest variance was in the largest or smallest group. When the largest variance was found in the largest group, the number of type I errors was significantly higher than what was expected under normal theory. When the largest variance was found in the smallest group, however, the number of type I errors was significantly less than what would be expected under normal theory.

When unequal group slopes were coupled with heterogeneous variances a different pattern emerged. When the largest variance was found in the smallest group, significant differences (at the $p < .01$ level) emerged; raw differences that were much higher than when the largest variance was paired with the smallest group in the equal n simulation. When the largest variance was paired with the largest group size, however, no significant differences could be found. It should be mentioned at this point that the largest group correlation (slope) is found in the third treatment group for both of these situations. Apparently, the coupling of the largest variance with the largest group size and largest regression slope improves the fit between the empirical and theoretical sampling distributions, while the coupling of the largest variance with the smallest group size and the smallest regression slope increases the disparity between the empirical and theoretical sampling distributions.

Table 2

Maximum Differences Between the Empirical and Nominal Sampling Distributions for the ANCOVA Simulations

Largest to Smallest Group Variance Ratio	BALANCED DESIGNS												UNBALANCED DESIGNS						
	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	2:1	3:1	5:1
Normally Distributed Covariate With Equal Regression Slopes																			
	Sample Size = 15				Sample Size = 30				Sample Size = 45				Sample Sizes = 15, 30, 45						
Distributional Shape																			
Normal	9	-23	-49	-70	49	27	14	-7	11	16	-13	-26	31	67*	83¶	87¶	-86¶	-142¶	-186¶
Platykurtic	23	-16	-50	-65	51	37	15	-6	19	16	19	-25	37	72*	84¶	90¶	-80¶	-140¶	-181¶
Leptokurtic	13	-18	-33	-63	11	11	6	-14	8	-6	-10	-23	22	61*	75¶	78¶	-90¶	-129¶	-177¶
Moderate Skew	-19	-21	-48	-81	40	31	12	-11	20	20	18	-22	15	69*	97¶	90¶	-134¶	-221¶	-325¶
Mod. Skew & Platy.	8	-21	-49	-80	38	18	15	-12	16	12	11	-21	46	77¶	97¶	97¶	-84¶	-159¶	-197¶
Mod. Skew & Lepto.	13	-17	-42	-69	16	11	7	-25	-4	8	-14	-24	24	79¶	92¶	75¶	-92¶	-140¶	-190¶
Extreme Skew & Lepto.	11	-24	-58	-95*	24	11	-13	-37	17	-11	-13	-25	32	50	68*	75¶	-96¶	-140¶	-207¶
Normally Distributed Covariate With Unequal Regression Slopes																			
Distributional Shape																			
Normal	-23	-47	-67	-81	-29	-33	-36	-52	22	30	21	8	-191¶	-9	19	31	-182¶	-221¶	-261¶
Platykurtic	-26	-42	-62	-79	-24	-29	-33	-42	30	32	32	23	-179¶	-14	24	34	-170¶	-206¶	-253¶
Leptokurtic	-20	-40	-52	-71	-37	-41	-49	-51	24	10	-10	-15	-177¶	-11	22	44	-182¶	-220¶	-282¶
Moderate Skew	-26	-44	-60	-84	-20	-30	-39	-41	25	29	18	-13	-165¶	-8	30	41	-171¶	-229¶	-267¶
Mod. Skew & Platy.	-21	-36	-53	-71	-14	-32	-34	-35	20	40	29	15	-141¶	9	24	51	-160¶	-216¶	-251¶
Mod. Skew & Lepto.	-30	-40	-56	-70	-22	-30	-38	-40	24	16	11	-13	-165¶	-5	23	48	-176¶	-222¶	-278¶
Extreme Skew & Lepto	-32	-35	-64	-102*	-27	-26	-35	-51	27	20	13	14	-170¶	7	30	44	-168¶	-208¶	-271¶

* Significant at the $p < .05$ level

¶ Significant at the $p < .01$ level

Effects of Assumption Violations on the Analysis of Covariance Using a Skewed Concomitant

Differences between the empirical and nominal F sampling distributions for the ANCOVA simulations using a skewed covariate are found in Table 3. For balanced designs involving small groups (n = 15) and a skewed covariate, no significant differences emerged. In fact those (statistically nonsignificant) differences that did emerge tended to be smaller in magnitude than those found when the same dependent vectors were used with normal covariates. The same can be said for the balanced designs using groups of 30 and 45.

When the unbalanced design was coupled with equal slopes and homogeneity of variances, no significant differences emerged. When the unbalanced design was coupled with heterogeneous slopes and homogeneity of variances, however, differences significant at the $p < .01$ level did emerge.

When heterogeneity of variance was coupled with equal regression slopes and unequal group sizes, the patterns identified originally with use of a normal covariate emerged again. Significantly less type I errors emerged when the largest variance was found in the largest group. However, when the largest variance was paired with the smallest group, there was a significant increase in the number of type I errors made.

When heterogeneity of variances was coupled with heterogeneous slopes and unequal n's, patterns emerged which were similar to those identified when the normal covariate was coupled with unequal slopes. When the largest variance was found in the smallest group, significant differences (at the $p < .01$ level) emerged; raw differences which were much higher than when the largest variance was paired with the smallest group in the equal slope situation. When the largest variance was paired with the largest group size, however, no significant differences could be found. Again here, like the analyses involving a normal covariate, the smallest correlation coefficient was found in the group with the smallest size. And again the coupling of the largest variance with the largest group size and largest regression slope improves the fit between the empirical and theoretical sampling distributions, while the coupling of the largest variance with the smallest size and the smallest slope increases the disparity. Once again, it is interesting to note that in many cases, use of the skewed covariate seemed to improve the fit between the empirical and nominal F sampling distributions.

FINDINGS AND CONCLUSIONS

Balanced Designs

Previous research (Glass, Peckham and Sanders, 1972; Harwell, Hays, Olds, and Rubinstein, 1990; etc.) suggest that heterogeneity of variances is the greatest single threat to robustness. Conventional thought suggests that when a balanced ANOVA or ANCOVA is used,

Table 3

Maximum Differences Between the Empirical and Nominal Sampling Distributions for the ANCOVA Simulations

Largest to Smallest Group Variance Ratios	BALANCED DESIGNS												UNBALANCED DESIGNS						
	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	2:1	3:1	5:1
Skewed Covariate With Equal Regression Slopes																			
	Sample Size = 15				Sample Size = 30				Sample Size = 45				Sample Sizes = 15, 30, 45						
Distributional Shape																			
Normal	±4	13	-18	-28	43	28	18	-17	17	7	-9	-21	27	83¶	100¶	103¶	-106¶	-152¶	-194¶
Platykurtic	8	16	-22	-27	48	28	23	-11	17	7	3	-17	30	87¶	94¶	102¶	-97¶	-194¶	-193¶
Leptokurtic	11	-2	-12	-20	22	14	8	-13	18	±6	-8	-19	18	72*	85¶	93¶	-98¶	-142¶	-190¶
Moderate Skew	6	-9	-25	-39	25	25	9	-19	9	12	-18	-21	30	83¶	98¶	101¶	-99¶	-151¶	-290¶
Mod. Skew & Platy.	7	-15	-27	-45	42	28	32	-10	17	8	-9	-19	55	88¶	101¶	103¶	-100¶	-198¶	-203¶
Mod. Skew & Lepto.	13	-9	-13	-21	23	7	-8	-12	5	15	-8	-20	14	68*	85¶	98¶	-94¶	-140¶	-192¶
Extreme Skew & Lepto.	7	-15	-17	-47	18	-10	-14	-28	15	10	-15	-24	30	78¶	90¶	94¶	-95¶	-141¶	-191¶
Skewed Covariate With Unequal Regression Slopes																			
Distributional Shape																			
Normal	-17	-21	-26	-40	-25	-27	-35	-45	14	9	10	-8	-166¶	-5	21	46	-169¶	-200¶	-249¶
Platykurtic	-13	-21	-22	-29	-18	-27	-30	-38	13	16	19	-7	-170¶	-10	31	31	-163¶	-200¶	-249¶
Leptokurtic	-13	-16	-24	-31	-25	-34	-35	-45	17	9	-4	-18	-154¶	-6	31	50	-154¶	-206¶	-249¶
Moderate Skew	-28	-24	-34	-49	-22	-28	-33	-42	19	11	3	-8	-158¶	-8	15	33	-152¶	-202¶	-255¶
Mod. Skew & Platy.	-18	-22	-26	-49	-13	-26	-33	-38	27	9	10	18	-166¶	-5	25	35	-158¶	-210¶	-245¶
Mod. Skew & Lepto.	-23	-33	-34	-38	-24	-37	-38	-43	21	11	7	-14	-162¶	-8	23	39	-164¶	-218¶	-262¶
Extreme Skew & Lepto	-34	-29	-41	-55	-30	-34	-43	-52	21	11	7	-16	-162¶	±10	21	37	-175¶	-219¶	-267¶

* Significant at the $p < .05$ level

¶ Significant at the $p < .01$ level

problems arise only when the ratio of largest to smallest group variance exceeds three. Meta-analytic findings by Harwell et al., however, suggest differently: balanced designs may suffer from inflated type I error rates when the ratio is as small as two.

The group variance ratios used in this simulation were chosen to directly compare Harwell et al.'s claim against the standard set by Glass et. al. No support was found for Harwell's claim; quite the contrary, there were almost no significant differences found in any of the balanced designs, even when the ratio between the largest and smallest group variance was as high as 5.

The results of this simulation when using a balanced design ANOVA or ANCOVA suggest a robustness far beyond that suggested by Glass et al. The unique methodology employed in this study may help to explain why. As part of the data generating process, the base vectors that had skew or kurtosis values significantly different from zero were systematically discarded, and new ones created. This procedure reduced the probability that the perturbations were a shape different than proported. Following removal of this sampling noise, the causes for the differences that remain are easier to isolate and interpret. Most of the studies that Glass et al. reviewed, however, used a methodology whereby parent populations with the desired characteristics were created and repeated random samples were drawn. No check was made to insure that the samples drawn possessed the mathematical properties being tested. Therefore, when significant differences emerged between the empirical and theoretical F distributions, it was unclear to what degree the differences were the result of the known mathematical characteristics and at what point they became the product of selected samples that, by the luck of the draw, possess mathematical properties far different from their parent populations.

The fact that the few significant differences that did arise in the balanced designs did so among the small group size (n = 15) is also worth noting. The confidence bands, used to screen out samples with mathematical characteristics different from those to be tested, are widest when the sample size is small. It is possible that some samples that should have been disgarded were not because of the wide confidence bands. If this is the case, then the origin of the significant differences that emerged in the small sample size simulations remains unclear: are they the result of violations of the assumptions under test, or are they the result of inclusion of extreme samples with mathematical characteristics different from those being tested?

Games and Lucas (1966) suggested that a skewed dependent is a greater threat to robustness than a leptokurtic or platykurtic dependent variable. Additionally, they have suggested that the validity of the F test improves for leptokurtic distributions but suffers when using platykurtic distributions. Distributional shape, however, did not prove to be a major factor in influencing type I error rates in this simulation.

Potthoff (1965) suggests that a non-normal concomitant increases the sensitivity of F to departures from normality in the dependent variable. This research found just the opposite: the

small (but statistically nonsignifiant) differences that did emerge found analyses using the normal covariate - not the skewed - to be most sensitive to distortions in the dependent variable.

Unbalanced Designs

Whereas the balanced design turned out to be very robust, the same cannot be said of the unbalanced design. Statistically significant differences emerged in face of almost all conditions except some that involved only perturbations of shape. Previous research (eg. Scheffe', 1959; Shields, 1976) have suggested that when heterogeneity of variance is coupled with unequal n's, the effect of the violation of equal variances will differ in nature depending on whether the larger group is paired with the larger or smaller variance. Specifically, they suggest that inflated type I error rates result when there is an inverse relationship between the group size and its variance, while deflated type I error rates will result when the larger group is paired with the larger variance.

Glass et al. (1972) suggest that the effects of nonnormal shapes and heterogeneous variances appear to be additive, something that this research supports. The idea of additive effects seems to extend beyond the match between distributional shape and heterogeneous variances, however. For instance, in the unequal n situation the smallest regression slope is paired with the smallest group size for all analyses. When this combination (which should increase the number of type I errors made) occurs jointly with heterogeneous variances where the smallest variance is found in the smallest group (which should decrease the number of type I errors), the net effect is a wash out; that is, no significant differences remain. Conversely, when the combination of the smallest slope and group size is paired with the largest variance, the number of type I errors increased dramatically - higher than either one of the violating conditions alone could have produced.

Concluding Remarks

In summary, for balanced designs the ANOVA and ANCOVA F statistics were found to be remarkably robust when faced with most of the violations included in this simulation. The degree to which the F test was robust, however, was surprising. The procedure remained robust even when the ratio of largest to smallest variance was as high as five. After the systematic removal of sampling noise due to the chance creation of skewed and/or kurtotic base vectors, F was found to be far more robust than previously believed. This research, however, reaffirms once again the findings of many previous studies that suggest that ANOVA and ANCOVA be avoided when group sizes are not equal.

In terms of specific recommendations to research practitioners using balanced designs, the ratio of largest to smallest group variance should continue to be checked. If the ratio is less

than 3, then the researcher need not fear invalid results due to any of the data set violations included here. If the ratio is between 3 and 5, however, the researcher should test to see if his or her dependent data is within the 95% confidence bands surrounding zero skew and kurtosis. If the dependent's skew and kurtosis values are within this range, then the F statistic should still be sufficiently robust. If, however, either the skew or kurtotic values fall outside of the 95% confidence band, then the researcher should consider the use of a statistical procedure with less stringent assumptions.

In terms of the direction of future research, several questions remain unanswered concerning the specific findings of this simulation. First, if the balanced designs (for group n's of 30 and above) are sufficiently robust when the largest to smallest group variance ratio is as high as five, then how high can that ratio get before robustness is significantly affected? Second, for equal sized samples smaller than size 30, are the confidence bands sufficiently narrow to provide researchers with the reassurance they need to use ANOVA or ANCOVA when the ratio of largest to smallest variance is between 3 and 5? Can use of smaller confidence bands (90% or 80% perhaps?) make up for the smaller sample size? Finally, this research used extremely unequal group sizes in the unbalanced designs (a difference of 300% between the largest and smallest groups). What would happen if the difference between the largest and smallest groups were smaller? How different can group sizes become before the robustness of the F statistic is jeopardized?

Finally, it should be noted that this research deals only with robustness. Robustness, however, is only the first of two issues that a researcher must consider when choosing a statistical procedure to analyze his or her data. The second issue involves power, and ultimately reduces to the following question first suggested in 1959 by Scheffe': which procedure from among those available will produce the most statistically accurate results in a specific research situation? It is in this direction that future Monte Carlo research of this genre must direct its attention.

BIBLIOGRAPHY

Atiquallah, M. (1964). The robustness of the covariance analysis of a one-way classification. Biometrika, 51, 365-372.

Bliss, C.I. (1967). Statistics in Biology. New York: McGraw - Hill.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the oneway classification. Annals of Mathematical Statistics, 25, 290 - 302.

Box, G.E.P. & Anderson, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, 17, 1 - 26, June, 1955.

Cochran, W.G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 3, 22 - 38.

Cochran, W.G. (1957). Analysis of covariance: Its nature and uses. Biometrics, 13, 261 - 281.

Elashoff, J.D. (1969). Analysis of covariance: A delicate instrument. American Educational Research Journal, 6, 383 - 401.

Evans, S.H. & Anastasio, E.J. (1968). Misuse of analysis of covariance when treatment effects and covariate are confounded. Psychological Bulletin, 69, 225 - 234.

Feldt, L.S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. Psychometrika, 23, 335 - 353.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. Psychometrika, 43, 521-532.

Games, P.A. & Lucas, P.A. (1966). Power of the analysis of variance of independent groups on nonnormal and normally transformed data.

Gibbons, J.D. & Chakrabort, S. (1992). Nonparametric Statistical Inference, 3'rd Ed. New York: Marcel Dekker.

Glass, G.V., Peckham, P.D. & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42 (3), 237 - 288.

Halpin, G. & Halpin, G. (1988). Evaluation of research and statistical methodologies. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY, November.

Hammersley, J.M. & Handscomb, D.C. (1967). Monte Carlo methods. London: Methuen & Co.

Harwell, M.R., Hayes, W.S., Olds, C.C. & Rubinstein, E.N. (1990). Summarizing Monte Carlo results in methodological research: the oneway, fixed-effects ANOVA case. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Lord, F. M. & Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison - Wesley.

McClaren, V.R. (1972). An investigation of the effects of violating the assumption of homogeneity of regression slopes in the analysis of covariance model upon the F-statistic. Unpublished doctoral dissertation, North State Texas University, Denton, TX.

McLean, J.E. (1974). An empirical examination of analysis of covariance with and without Porter's adjustment for a fallible covariate. Unpublished doctoral dissertation, University of Florida.

McLean, J.E. (1979). The Care and Feeding of ANCOVA. Paper presented at the annual meeting of the Mid-South Educational Research Association in Little Rock, AR, November.

McLean, J.E. (1989). ANCOVA: A Review, Update and Extension. Paper presented at the annual meeting of the Mid-South Educational Research Association in Little Rock, November.

Pearson, E.S. (1929). The distribution of frequency constants in small samples from nonnormal symmetrical and skew populations. Biometrika, 19, 151 - 164.

Peckham, P.D. (1968). An investigation of the effects of non-homogeneity of regression slopes upon analysis of covariance. Unpublished doctoral dissertation, University of Colorado.

Potthoff, R.F. (1965). Some Scheffe´ type tests for some Behrens-Fisher type regression problems. Journal of the American Statistical Association, 60, 1163 - 1190.

Raaijmakkers, J.G. & Pieters, J.P.M. (1987). Measurement error and ANCOVA: Functional and structural relationships. Psychometrika, 52, 4, 521 - 538.

Scheffe´, H. (1959). The analysis of variance. New York: John Wiley & Sons.

Shields, J.L. (1978). An empirical investigation of the effect of heteroscedascity and heterogeneity of variance on te analysis of covariance and the Johnson-Neyman technique. Army Technical Paper 292, Project no. 2Q762722A777, July.

Student, (W.S. Gossett). (1908). The probable error of the mean. Biometrika, 6, 1 - 25.

Winer, B.J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.